基于词语相关性的对话系统话题分割 *

何天文 [1, 2]，王　红 [1, 2, 3†]，刘海燕 [1, 2]

(1. 山东师范大学 信息科学与工程学院，济南 250358; 2. 山东省分布式计算软件新技术重点实验室，济南 250014; 3. 山东师范大学 生命科学研究院，济南 250014)

摘　要：针对开放域对话系统中存在的话题转移问题以及对话内容中存在大量短文本的情况，传统的基于相似性的处理方法存在很大的局限性，创新地提出通过对话系统中前后句子的相关性判断分割点，实现话题分割，并比较了相关性与相似性在计算中对句子信息利用的不同之处。提出一种相关性计算方法，并将该方法用于话题分割，最终实现话题转移检测。通过与现有方法的对比实验，表明了提出的相关性计算方法的有效性。

关键词：相关性；话题分割；话题转移；对话系统

中图分类号：TP391.1

Topic segmentation of dialogue system based on correlation of words

He Tianwen[1, 2], Wang Hong[1, 2, 3?], Liu Haiyan[1, 2]

(1. *School of Information Science & Engineering Shandong Normal University*, *Jinan* 250358, *China*; 2. *Shandong Provincial Key Laboratory for Distributed Computer Software Novel Technology*, *Jinan* 250014, *China*; 3. *Institute of Biomedical Sciences Shandong Normal University*, *Jinan* 250014)

Abstract: In view of the problems of topic transfer and the existence of a large number of short text in the dialogue content in open domain dialogue systems, the traditional similarity-based processing method has many limitations. This paper proposeed an innovative method, which is based on the relevance of the sentences to determine whether the dialogue topic transfer, and compares the difference between the correlation-based and the similarity-based methods in revealing the relationship between sentences. Furthermore, this paper presents a correlation-based algorithm to calculate the correlation of words and apply it to segment topics of sentences, and this can address some challenges of topic transfer detection. Comparing with existing methods, the experimental results demonstrate the superior performance of the correlation-based method in this paper.

Key Words: correlations; topic segmentation; topic transfer; dialogue system

0　引言

随着计算机技术与大数据产业的快速发展，人机对话系统亦得到迅猛发展，进而推动了对话系统的研究如火如荼地展开，在学术界和工业界受到广泛关注。目前，人机对话系统的研究成为人工智能领域一项非常重要而且极具挑战性的工作[1]，研究中存在许多亟待解决的问题。

对话系统的核心任务就是根据历史对话信息生成应答语句[2]。有效完成该任务的关键是话题追踪。话题追踪负责检测整个对话过程中的话题转变[3]，实现话题分割，在系统生成应答语句过程中能够根据当前话题生成话题相关语句或话题引导语句，使对话系统不会出现“所答非所问”，这正是本文要解决的问题。

显然，话题分割的依据是对话系统中聊天内容，它为查找和生成应答语句提供非常重要参考。但是，这些历史对话语料信息有其自身的特殊性，比如：聊天语句有可能会很短，聊天语句中的指代现象过于严重，等等。当聊天语句很短的时候，如“好的”，只能从句子中获得一种肯定的态度信息，其他信息则很难获得。当聊天语句中指代现象太过严重时，如“他给她打了个电话”“已经康复了”等只看单个句子根本无法理解其准确意思，而这种指代和省略现象在口语中非常常见，任何一种语言的口语中都存在的问题[4]。另外，对话系统中用户给出的句子可能不符合标准的语言规范，这种情况增加了对话系统中通过规则和模板处理语句的难度。目前，依据聊天语料进行话

基金项目：国家自然科学基金资助项目（61672329，61373149，61472233，61572300，81273704）；山东省科技计划资助项目（2014GGX101026）；山东省教育科学规划项目（ZK1437B010）；山东省泰山学者基金项目（TSHW201502038，20110819）；山东省精品课程项目（2012BK294，2013BK399，2013BK402）

作者简介：何天文（1992-），男，山东济南人，硕士研究生，主要研究方向为自然语言处理、机器学习、数据挖掘（sdsfhtw@163.com）；王红（1966-），女（通信作者），天津人，博士，主要研究方向为移动社会软件、复杂网络、工作流；刘海燕（1995-），女，山东济宁人，硕士研究生，主要研究方向为机器学习、数据挖掘．

题分割的工作是基于相似度、边界和概率图模型等算法实现的，计算过程中通过计算文本句子上下文之间的相似度来判定话题的边界。而忽略了上下文的相关性关系。文本相似度虽然能在一定程度上计算出上下文的语义关系，但是文本前后句子的关系除了相似关系还有上下位分等级的关系和相关关系。

针对上述问题，本文完成如下工作：

a)分析了相似度局限性，针对对话系统话题分割的划分分割点的要求，提出了一种在语义空间中构建词向量的哈夫曼编码树，计算词语相关性的方法；b)将词语的相关性计算信息拓展到对话系统中句子的相关性计算；c)根据句子之间的相关性识别对话系统上下文中不同话题的边界，由此判断话题转移概率并进行话题分割实现话题转移检测。

1　相关工作

1.1　相似性的局限性分析

在自然语言处理领域，将文本文档作为聊天语料进行分割研究的工作已经非常广泛，现有工作分别基于相似度、边界和概率图模型等提出了一些方法，但是这些方法中大多采用计算文本句子上下文之间的相似度来判定话题的边界，文本相似度虽然能在一定程度上计算出上下文的语义关系，但是文本前后句子的关系除了相似关系还有上下位分等级的关系和相关关系。刘群等为了区分语义相似与语义相关给出了相关性的概念定义[5]。相似性量化并不是面向相关关系的，概念不一致[6]。相似性表示词汇具有某种可替代性，有某些相同内涵特征或者属性特征；相关性表示词汇语义上具有某种相互依赖、相互影响的特征。因而，在聊天语料话题分割任务中，确定话题分割点时以句子内容之间的相关性作为判断依据，比以相似性做判断，会有更高的准确性和合理性。基于以上对文本相似性和相关性的分析，本文采用词汇相关度计算代替其他模型使用的词汇相似度计算。

1.2　相关性研究

在上一部分中已经对文本的相关性给出解释。相关性的计算中，信息熵[7]可以用来表示词语关系的不确定性程度，单个词语 x 熵的计算如式(1)所示。

$$S=-\sum_{x}P(x)\log P(x) \tag{1}$$

其中：$P(x)$ 表示词语 x 出现的概率。而计算词语 x 与词语 y 的信息熵则如式(2)所示。

$$S(Y\,|\,X)=S\{P(x,y)\}-S\{P(x)\} \tag{2}$$

其中要计算在已知词语 x 情况下再获得词语 y 的信息熵，其中 $P(x,y)$ 则用来计算在开放域语料上两个词语的共现概率，在后面计算相关性分析中借鉴了这种思想。另外的协方差和 SVD 计算方法都存在计算量大和语料规模需求量大的问题，不太适合用于开放域文本的相关性计算。

谌志群等人[8]利用中文维基百科数据集中的分类体系和页面关联链接等实体和关系信息计算词语的相关度。页面距离计算方法是检索距离法，即在检索结果中的相似度。具体方法是通过计算斯皮尔曼等级相关系数判断检索结果中的数据是否具有一致性，一致性程度通过 Consine 距离计算得出。该方法使用的是数据的相似度做计算，另外由于开放域对话系统的实体数和关系数量众多，使用维基百科数据集存在一定有限性。

Song 等人在 TextTiling 基础上提出了启发式方案计算相似度。TextTiling 中的相似度计算如式(3)所示。

$$\mathrm{sim}(S_1,S_2)=\cos(s_1,s_2)=\frac{s_1^{\mathrm{T}}s_2}{\|s_1\|\cdot\|s_2\|} \tag{3}$$

改进后的计算方法如式(4)所示，

$$\mathrm{sim}(S_1,S_2)=\frac{1}{n_1}\sum_{i=1}^{n_1}\max_{j=0}^{n_2}\left\{\cos\left(w_i,v_j\right)\right\} \tag{4}$$

其中：w_i 和 v_j 分别是 S_1 和 S_2 中的一个词语，n_1 和 n_2 分别表示 S_1 和 S_2 的词语数。即将逐个词语计算所有词对相似度修改为将一组比较中最高相似度记为句子的相似度。本文的工作将在此向量空间和相似度计算方法基础之上改用相关度计算来确定对话系统上下文中的话题边界，进行对话分割。

1.3　话题模型

在自然语言研究领域，已经有一些用来处理语义空间概念和话题的文本的主题话题模型，如 LDA、LSI 和 LSA 等。其中 LSI 通过在数据集上进行矩阵分解构建语义空间，再将文档内容和词语映射到语义空间进行计算。这些方法通过词语共现的关联关系表示语义空间，这种情况下，某个词语可能和一个话题相关性特别强，但也会因为语料规模和质量的关系而表现出较弱的相关性[9]。这些基于词频共现和矩阵分解的方法忽略了词序列顺序，用于短文本较多的口语对话系统，会降低算法在文本文档上表现的性能。

1.4　对话分割

在文本分割的相关研究中，最早由 Hearst[10]提出 TextTiling 分割方法。先将文本划分为句子级单位，再对上下文中的各个单位的关联性进行打分。根据 Cosine 距离来计算各单位相似性，再按照相似性划分话题边界。

Liu 等人[11]的文本分割方法与 TextTiling 方法思路相近，分割边界的评判用的是词语在语义空间中的 Cosine 距离，之后同样使用经验阈值做分割。将这部分工作与 LSA 结合用于为中文文本生成摘要。

邹博伟等人[12]降低了 TextTiling 方法中模型对下文内容的依赖，用相对坡度下降值代替传统方法中的绝对坡度下降值，这样可以有效地解决连续 query 之间的相似度低造成的段落无法正确划分的问题。改进后的句子间相对深度计算式(5)所示。

$$\begin{aligned}depth_i=&\max\left\{\frac{sim_{i-1,i}-sim_{i,i+1}}{sim_{i-1,i}},0\right\}+\\&\max\left\{\frac{sim_{i+1,i+2}-sim_{i,i+1}}{sim_{i+1,i+2}},0\right\}\end{aligned} \tag{5}$$

其中：$depth_i$ 表示相邻句子相似度的相对坡度，$sim_{i,i+1}$ 表示当前句子与下一句的相似度。该改进方法将性能提高了 3.8%。

Joty 等人[13]对 TextTling 算法进行了改进，不再使用阈值界限作为分割标准，而是采用自上而下的分层聚类方法，将句子序列按照不同话题进行切分，取得了较好的效果。Malioutov 等人[14]则将句子的分割问题转换为图分割问题，并且在 TF-IDF 特征基础上提出了一种最小分割模型，通过切分演讲报告，验证了方法的有效性。Ye 等人[15]通过改进 Dotplotting 算法，最小化两类话题间相关性，同时最大化两类话题内相关性，以此更精确地划分不同话题，从而获得较好的文本切分效果。

人类语言的结构一直都没有完好的结构或规则，人们组织的口语对话句子中经常出现冗余、偏差和不符合书面语表达的情况，所以有学者提出对规范化文本进行话题分析。Blei 等人[16]已提出针对在 Science 杂志上发表的文章的相关性分析和研究，指出了这类文章的编排有固定的格式，而且杂志上发表的文章里面的句子一般都符合标准的语义语法结构，符合人们共同认同的书面表达习惯。作者分析了各个文章的相关性，并提出了针对这种科技型文章的话题模型，实验证明了模型的有效性。

童毅见等人[17]在自动文摘任务中使用主题划分，其方法融合了特定语言现象和文本特征，例如二元词组频率和命名实体重复等情况。取得了不错的效果，但是因开放域对话的语境的特性，无法将该方法推广到开放域对话系统中。

El-Kishky 等人[18]采用高频短语分割文档，考虑到语言的非组合性原则，先采用词组挖掘方法分析出高频重要短语，再用这些高频短语结合统计方法分割文本，同时还要过滤一些级联短语。在候选话题分配过程中将短语各部分约束到一个共享的候选主题下，最终确定主题分布。

Song 等人改进 TextTiling 算法时提出相邻句子中的重复信息具有的计算价值，并修改了词汇向量空间的生成方法，提出了“virtual sentences”映射到向量空间做相似度计算。很多的文本分割方法并不能直接用于对话系统中的话题分割，因为对话中可用的信息只有上文，而文本文档中是上下文全部的信息，但是上面提到的这些分析方法和方案都具有借鉴意义，本文将开放域文本的相关度计算应用于对话系统的话题分割，以通过文本相似度之外的其他隐含语义判定话题边界，提高话题分割的准确率。

2 对话系统话题分割方法

2.1 词语相关性计算

在之前的工作中，很多分割模型话题边界的确定依赖于对上下文句子相似性的判断，本文将改用计算上下文句子的相关性来判断话题的边界。

相关性表示两个词的互相关联程度，即从一个词关联到另一个词的概率，也可表示两个词出现在同一句话或相邻两句话中的概率，计算公式如式(6)所示。

$$Correlation(w_i,w_j)=P(w_j|w_i) \tag{6}$$

其中：w_i 和 w_j 表示需要计算相关性的两个词语。因为需要在构建词向量时既要在向量中包含词语在句子中的位置信息，还需要可以快速索引，因此在普通的 one-hot 模型空间中是无法计算的。本文借鉴 Google 的 Word2Vec 模型[19]中 Skip-Gram 的思路训练词向量。在训练好的语料的向量空间基础上，结合 Huffman Softmax 模型中的编码信息与目标词对应的词向量，最终计算得出两个词相关的似然概率[20]。将在网络上抓取的大规模文本数据分词处理后使用 Word2Vec 模型训练得到语料中词语的词向量，训练过程中还产生了以这些词向量作为叶子节点的 Huffman 编码树，其中的非叶子节点中存储中间向量，这些向量代表了它对应的所有子节点，即可以通过计算得出目标词向量与当前节点下叶子节点中向量的条件概率。先获取词语 w_j 的 Huffman 编码路径序列 C，将需要计算相关性的词语 w_i 在向量空间中的词向量与编码路径 C 上的各个节点 c 计算得出整个路径上的预测概率，各个节点的概率计算过程如式(7)所示。

$$Pl(i,\theta,c)=\log\left(\frac{1}{1+e^{-i^{T}\theta}}\right)^{1-c}\left(1-\frac{1}{1+e^{-i^{T}\theta}}\right)^{c} \tag{7}$$

其中：i 表示输入的词向量，θ 表示节点向量，其中 $c\in\mathrm{C}$，表示源向量到目标词向量路径上节点的编码序列，求得预测目标词语向量过程中在各个节点的概率，再将整条路径上算出的概率相乘，最终得到两个词语的似然概率，计算过程如式(8)所示。

$$P(w_j|w_i)=\mathrm{P}(j|C,i)=\prod_{c}^{n}\alpha\cdot Pl(i,\theta,c) \tag{8}$$

算法 1. 词语的相关性计算

```
Input:VectorSpace,wᵢ,wⱼ
Output:P
Correlation(VectorSpace,wᵢ,wⱼ)
  return P
wᵢ ← GetVector(VectorSpace,wᵢ)
Cj ← GetHPath(VectorSpace,wⱼ)
P ← 1.0
for all c∈Cⱼ do
  p ← α*Pl(wᵢ,θ,c)
  P ← P*p
end for
```

其中：i 表示词语 w_i 在向量空间中的向量，j 表示词语 w_j 的向量表示，α 是一个需要训练的超参数，表示路径上的距离惩罚系数，用于平衡不同距离上的词语对预测概率的影响，具体计算过程如算法 1 所示。通过该算法计算得到两个词语的似然概率，代表这两个词语同时出现的概率，即两个词语的相关性。

2.2 对话系统话题分割

假设相邻两个句子分别为 S_1 和 S_2，首先根据 TF-IDF 算法和规则过滤获取句子的关键词[21]，根据 Wu 等人[22]处理匹配应答语句时的参数，分别取每句话相应的关键词进行相关性(correlation)计算。每个句子提取核心关键词作为其话题关键词，用于确定句子的话题和句子中所含具体内容。将 S_2 中的每一个

词语与 S_1 中的所有词语进行相关性计算，考虑到口语对话中句子内容的随意性和复杂性，为了降低不规范文本对概率计算的影响，取一个词对应的相关性最大值作为该词与句子 S_1 的相关性。两个句子的相关性则用各个词的相关性均值表示。其相关性计算公式如式(9)所示。

$$Corr(S_2 \mid S_1) = \frac{1}{n}\sum_{j=0}^{n}\max_{i=0}^{m}\left\{P\left(w_j \mid w_i\right)\right\} \tag{9}$$

其中：m 和 n 分别表示句子 S_1 与 S_2 分词后所包含词语的个数，w_i 与 w_j 分别代表句子 S_1 与 S_2 中的词，$\max(\cdot)$ 表示取集合中最大概率值。

为了将词语相关性计算融入对话系统的话题分割，本文借鉴 N-gram 滑窗和 TextTiling 算法中的阈值判定话题边界思想。将对话系统中的聊天记录整理为句子序列 $DT = \{S_1, S_2 \cdots S_n\}$，以句子对作为滑窗基本单位做句子相关性判断，模拟两个人在对话，再通过训练获得话题分割的阈值，计算公式如式(10)所示。

$$seg(S_i, S_{i+1}) = \begin{cases} 1 & if\ Corr(S_i, S_{i+1}) \geq \sigma \\ 0 & otherwise \end{cases} \tag{10}$$

其中：S_i 与 S_{i+1} 是句子序列 DT 中前后相邻两个句子，$Corr(\cdot)$ 用于计算两个句子的整体相关性，σ 为训练得到的分割阈值，$seg(\cdot)$ 计算两句话中间是否存在分割点。

如算法 2 所示，向分割检测函数输入连续待分割的句子之后就可以判断中间是否需要设置分割点。

算法 2. 对话系统话题分割

```
Input:Si,Si+1
Output: seg
1:Segment(Si,Si+1)
return seg
2:r ← 0.0  maxresult ← 0.0
3:for all wj∈Si+1 do
4:  for all wi∈Si do
5:    result[] ← Corr(wj|wi)
6:  end for
7:  maxresult ← max(result[]) + maxresult
8:end for
9:r ← maxresult/len(Si+1)
10:seg ← Seg(r)
```

3　实验分析

3.1　实验准备

实验中使用的词向量空间，是用多个领域的 800 万篇文章训练得到的，是中文的平衡语料库，其中还包含了常见英文词汇。使用 Google 的 Word2Vec 进行训练，向量维度为 256 维，训练时设置窗口大小为 10，最小词频限制为 64[23]。训练数据和测试数据则使用 Wu 等人的文章提供的公开聊天语料[24]，原数据中含有从微博及豆瓣讨论组爬取的多轮对话数据，数据的数据量、对话轮数等具体信息如表 1 所示。

表 1 多轮对话数据

项目	训练集	验证集	测试集
对话数	1m	50k	10k
对话的平均正样本数	1	1	1.18
Fless Kappa	N\A	N\A	0.41
对话中最小轮数	3	3	3
对话中最大轮数	98	91	45
每组对话的平均轮数	6.69	6.75	5.95
每组对话的平均词数	18.56	18.50	20.74

对开放的数据进行随机采样，最终使用 50 万组对话构成训练集，2.5 万组对话构成验证集，测试集中数据为 1000 组。数据集中已经人工对对话数据是否存在话题转移做了标注，数据样例如表 2 所示，标签就是对话对应的标注，黑体字表示可能出现话题转移的位置，标签为 1 的表示句子描述内容属于同一话题，标签为 0 的表示出现话题转移，前后内容不连贯。

表 2 训练数据集中的数据样例

标签	对话数据
1	昆明 那里 配 眼镜 比较 便宜/云大 附近 很多 店 应该有 竞争 价格 会 下来 一点 的 吧/给 推荐 个 云大 附近 的 吧 谢谢/**去 了 就能 看到 比如 云光 什么 的**
0	昆明 那里 配 眼镜 比较 便宜/云大 附近 很多 店 应该有 竞争 价格 会 下来 一点 的 吧/给 推荐 个 云大 附近 的 吧 谢谢/**你 的 他 毕竟 还是 说 了 我 的 完全没有 任何 消息 我 伤害 了 他 于是 15 天 没 消息**

3.2　对话系统中话题分割

为了验证相关性与相似性在话题分割这个具体应用场景中的差异性，本文设计了对话系统的话题分割实验。为了测试不同边界分割阈值对对话中话题分割准确率的影响，本文选择了在训练集上准确率较高的三个精确到个位的分割阈值在测试集上进行了测试，测试结果如图 1 所示。

从图中可以看出，在阈值为 24.0 时准确率达到 0.544，而另外的两个阈值的准确率则在 0.515 到 0.530 范围内波动。分割阈值需要判断所有对话是否出现了话题转移，数据集中存在部分话题转移不是特别明确的一些对话，造成所有阈值的准确率在某一部分数据集上准确率都有所下降，如在数据量为 400 和 800 时，准确率都上升说明对话中出现了明显的话题转移。图中显示出准确率最高的分割阈值波动频繁但趋于稳定，与其他阈值的准确率在整体趋势上保持一致性。

在对比实验中，本文选择的对比模型与 song 选择的对比模型一样：一个是随机分割模型，另一个是结合了 TF-IDF 的 TextTiling 模型，在对比实验中代表相似性方法。随机分割模型中，本文参考已有工作的处理方式，得分的获取并不是完全随机分割的处理方法，而是在随机过程中添加了部分先验知识。先验知识的作用是通过训练集中正样本的高频词对测试集句子

中的关键词进行约束。另一个对比模型则是使用相关实验中的对比模型，在经典 TextTiling 方法上做的改进是在分割计算过程中融合了上下句文本中词语的 TFIDF 特征信息。实验中使用三个方法在测试数据集上进行测试，实验的结果如图 2 所示。

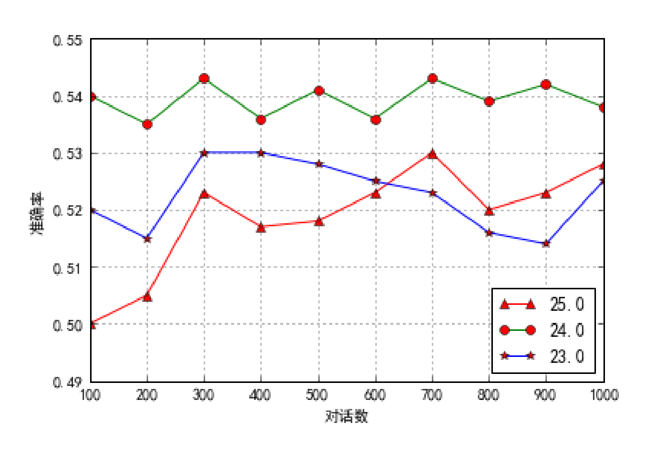


图 1　不同阈值在测试集上的准确率

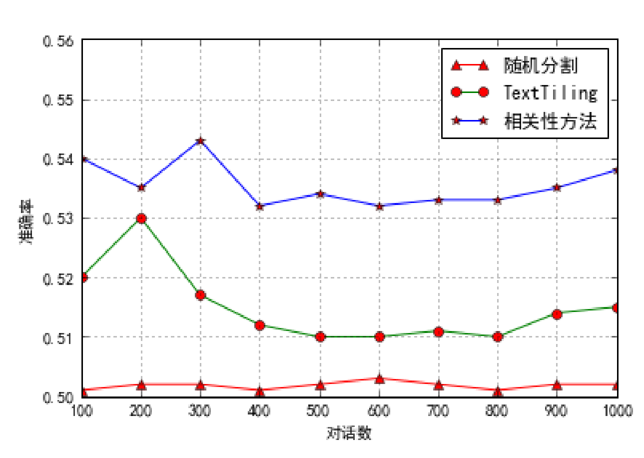


图 2　不同方法在测试集上的准确率对比

图中对比的是三个方法在测试集上的准确率，从图中可以看出，三个方法的准确率都高于 50%，说明随机分割方法中的先验知识在判断分割点时候也起到了一定的作用。而且相关性计算方法比利用文本相似性的 TextTiling 准确率高出 2%。随着数据量的增加 TextTiling 方法和相关性方法都有大幅波动，经过对实验数据分析认为这些波动是由数据中的较短文本造成的，虽然标注者能够判断出来，但是类似“怎么”、“谢谢”、“好的”、“可以”等极短的文本会对分割阈值判断分割边界准确性造成较大影响。图中显示出相关性方法的准确率比 TextTiling 方法的准确率高，相关性方法在处理相关性计算时采用了最大采样方法，以尽可能放大关键词之间的相关性，同时减小低频词、新词和专业词汇的相关性对句子整体相关性的影响，增加了方法的鲁棒性。TextTiling 算法中的所有词都会参与运算，结果中会包含多组相似性为 0 的结果，不能降低短词语对相似性计算准确度的影响，造成得到的句子之间的相似性会有较大的偏差。song 的模型在词向量映射过程中使用了改进后的 Word2Vector 方法，因缺少数据和程序而无法复现，但其文章中给出话题分割任务中的准确率为 0.521，而本文提出的相关性计算方法的准确率能达到 0.54。

本文还在比较不同方法准确率基础上，对方法的召回率进行了比较，召回率的比较结果如图 3 所示。图中显示了随机分割、TextTiling 和相关性计算三种方法的召回率，相关性方法的召回率也是最高的。

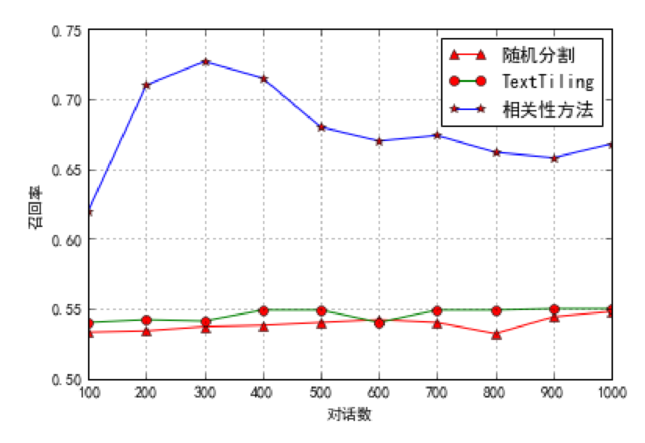


图 3　不同方法在测试集上召回率对比

从图 3 中可以看出 TextTiling 和相关性计算两种方法的召回率表现出小幅波动，而且相关性方法开始的波动较大，说明方法能识别出大部分需要分割的数据。这一部分也是分割阈值准确率波动到最高的数据段，说明受分割阈值的影响同时提高了准确率和召回率。但是当数据量达到 600 之后召回率基本趋于稳定，和准确率趋势是一样的。通过以上相关性方法与对比方法的比较，证明该方法的有效性。

4　结束语

本文首先分析了相似性与相关性在语义计算过程中的不同之处，分析了相似性计算存在的问题，并在已有工作基础上提出在词向量空间中计算词语相关性的方法。针对对话系统中话题分割任务，通过计算句子中词语的相关性，使用最大采样方法计算出句子之间的相关性。以此将计算词语相关性的方法拓展到计算对话系统中上下文句子的相关性，并将该方法用于在开放域范围内确定对话中话题分割点的位置。根据确定的分割点划分对话中的不同话题，实现话题转移检测。通过与其他分割方法的对比，实验结果显示准确率比对比方法提高了 2%，有较高的召回率，证明了相关性计算的有效性。

本文提出的利用相关性分析进行对话系统话题分割的方法还存在一些不足。词语语义相关性计算过程中对词语的向量空间依赖较大，因为对话系统中的分割任务面向的是开放域，向量空间的有限性也会降低一些新出现的词汇间相关性的准确率。用于判断话题转移设置分割点的相关性阈值的确定，受训练数据集的影响较大，而且该阈值的确定需要较大数据量，同时对数据质量要求较高。下一步工作将针对上述几个问题展开，进一步提高算法鲁棒性。

参考文献：

[1] Metallinou A, Bohus D, Williams J, et al. Discriminative state tracking for spoken dialog systems [C]// Proc of Meeting of the Association for Computational Linguistics. 2013.

[2] Song Yiping, Yan Rui, Li Xiang, et al. Two are better than one: an ensemble of retrieval-and generation-based dialog systems. 2016. arXiv: 1610. 07149.

[3] López-Cózar R, Callejas Z, Griol D, et al. Review of spoken dialogue systems [J]. Loquens, 2014, 1 (2): e012. 10. 3989//loquens. 2014. 012.

[4] Sikdar U K, Ekbal A, Saha S, et al. Differential evolution-based feature selection technique for anaphora resolution [J]. Soft Computing, 2015, 19 (8): 2149-2161.

[5] 刘群，李素建. 基于《知网》的词汇语义相似度计算 [J]. 中文计算语言学, 2002 (7): 59-76.

[6] 钟茂生，刘慧，刘磊. 词汇间语义相关关系量化计算方法 [J]. 中文信息学报, 2009, 23 (2): 115-122.

[7] 廖大麟. 随机事件的不确定性或信息量的度量——信息熵 [J]. 毕节学院学报：综合版, 2006, 24 (4): 35-38.

[8] 谌志群，高飞，曾智军，等. 基于中文维基百科的词语相关度计算 [J]. 情报学报, 2012, 31 (12): 1265-1270.

[9] Pu X, Jin R, Wu G, et al. Topic modeling in semantic space with keywords [C]// Proc of Conference on Information and Knowledge Management. 2015: 1141-1150.

[10] Hearst M A. TextTiling: segmenting text into multi-paragraph subtopic passages [J]. Computational Linguistics, 1997, 23 (1): 33-64.

[11] Liu C, Wang Y, Zheng F, et al. Using LSA and text segmentation to improve automatic Chinese dialogue text summarization [J]. Journal of Zhejiang University Science, 2007, 8 (1): 79-87.

[12] 邹博伟，张宇，范基礼，等. 基于改进 TextTiling 方法的用户新兴趣发现的研究 [J]. 计算机研究与发展, 2009, 46 (9): 1594-1600.

[13] Joty S, Carenini G, Ng R T. Topic segmentation and labeling in asynchronous conversations [J]. Journal of Artificial Intelligence Research, 2014, 47 (1): 521-573.

[14] Malioutov I, Barzilay R. Minimum cut model for spoken lecture segmentation [C]// Proc of International Conference on Computational Linguistics and Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, 2006: 25-32.

[15] Ye N, Zhu J, Wang H, et al. An Improved Model of Dotplotting for Text Segmentation. [J]. Journal of Chinese Language and Computing, 2007.

[16] Blei D M, Lafferty J. A correlated topic model of Science [J]. The Annals of Applied Statistics, 2007, 1 (1): 17-35.

[17] 童毅见，唐慧丰. 面向自动文摘的主题划分方法 [J]. 北京大学学报：自然科学版, 2013, 49 (1): 39-44.

[18] El-Kishky A, Song Y, Wang C, et al. Scalable topical phrase mining from text corpora [J]. Proceedings of the VLDB Endowment, 2014, 8 (3): 305-316.

[19] Mikolov T, Chen K, Corrado G, et al. Efficient estimation of word representations in vector space [J]. Computer Science, 2013.

[20] Mnih A, Hinton G. Three new graphical models for statistical language modelling [C]// Proc of International Conference on Machine Learning. ACM, 2007: 641-648.

[21] 牛萍，黄德根. TF-IDF 与规则相结合的中文关键词自动抽取研究 [J]. 小型微型计算机系统, 2016, 37 (4): 711-715.

[22] Wu Yu, Wu Wei, Chen Xing, , et al. Sequential matching network: a new architecture for multi-turn response selection in retrieval-based chatbots [C]// Proc of the 55th Annual Meeting of the Association for Computational Linguistics. 2016: 496-505.

[23] Su Jianlin. Incredible Word2Vec [EB/OL]. [2017-05-08]. http://spaces. ac. cn/archives/4304/,

[24] Wu, Yu, et al. "Sequential Matching Network: A New Archtechture for Multi-turn Response Selection in Retrieval-based Chatbots. "ACL. 2017.